
Inteligencia Artificial

Clasificación de Texto utilizando Naive Bayes

Jorge Luis Guevara Diaz

www.jorge.sistemasyservidores.com

Clasificación de Texto : Ejemplos

- Clasificar noticias como, negocios, ciencia, deporte, etc
- Clasificar emails como spam y no spam
- Clasificar archivos pdf
- Clasificar criticas sobre peliculas como bueno malo regular, etc
- Clasificar papers tecnicos, como interesante, no interesante, etc
- Clasificar sitios web de companias

Algoritmo de Aprendizaje Probabilístico

Naive Bayes Multinomial

- Realizar

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- Luego utilizar algoritmo MAP

$$c_{map} = \arg \max_{c \in \mathcal{C}} \hat{P}(c|d) = \arg \max_{c \in \mathcal{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

$$c_{map} = \arg \max_{c \in \mathcal{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

Algoritmo de Aprendizaje Probabilístico

Naive Bayes Multinomial

- Para calcular la probabilidad de C usar algoritmo ML

$$\hat{P}(c) = \frac{N_c}{N}$$

- Para calcular las probabilidades condicionales

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

Algoritmo de Aprendizaje Probabilístico

Naive Bayes Multinomial

- Para evitar problemas con los ceros, agregar suavizado laplaciano

$$\hat{\theta}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

- Donde $B=|V|$ es el número de términos en el vocabulario

Ejemplo

	Id doc	palabras	En c = Peru
Conjunto de entrenamiento	1	Lima Beijing Lima	Si
	2	Lima Lima Shanghai	Si
	3	Lima Macao	Si
	4	Tokyo Japon Lima	No
Conjunto de prueba	5	Lima Lima Lima Tokyo Japon	

-
- $P(\text{clase} | \text{documento})?$
 - $P(\text{clase} | \{\text{Lima Lima Lima Tokyo Japon}\})$

 - $P(\text{Lima} | c = \text{Peru}) = (5+1)/(8+6)$
 - $P(\text{Tokio} | c = \text{Peru}) = (0+1)/(8+6)$
 - $P(\text{Japon} | \text{Peru})?$
 - $P(\text{Lima} | c = \sim \text{Peru})?$
 - $P(\text{Tokio} | c = \sim \text{Peru})$
 - $P(\text{Japon} | \sim \text{Peru})?$
 - $P(c = \text{peru} | \{\text{Lima Lima Lima Tokyo Japon}\})$
 - $= 3/4 * 3/7^3 * 1/14 * 1/14 = 0.0003$

 - $P(c = \sim \text{peru} | \{\text{Lima Lima Lima Tokyo Japon}\})$
 - $1/4 * 2/9^3 * 2/9 = 0.0001$

Podemos concluir que el documento pertenece a la clase Peru

- Regla general

- $P(a|b) = \frac{\text{\# de ocurrencias de a dentro de b} + 1}{(\text{tamaño de b}) + \text{tamaño del vocabulario}}$